(MACHINE) LEARNING WHAT POLICYMAKERS VALUE

Anonymous author(s)

Abstract

When a decisionmaker allocates resources, the values behind the allocation are not always transparent. This paper develops a method to uncover the values implied by observed allocations. Our method relies on machine learning estimators to separately identify three components behind the allocation: (i) *implied welfare weights*: the decisionmaker may prioritize some people over others; (ii) *heterogeneous treatment effects*: some people may benefit more than others from the allocation; and (iii) *differential impact weights*: the decisionmaker may prioritize some outcomes over others. We apply this approach to Mexico's flagship anti-poverty program, to estimate the preferences its allocation implies. We find that the distribution of benefits are consistent with welfare weights that rank a household 13 percentiles higher if indigenous, 8 percentiles lower for each standard deviation increase in household income, and 21 percentiles higher for each additional small child in the household, on average. Allocations are consistent with valuing each missed school day and child sick day within conventional assessments. Alternate eligibility criteria could have improved average consumption, health or schooling outcomes.

1 Introduction

Decisions reflect values. Governments decide which households receive welfare benefits, which firms receive small business grants, and which regions receive aid; colleges decide which students to admit; journals decide what papers to publish. In each of these settings, an *allocation decision* is made. What values are implied by the observed allocation decision?

Decisions may prioritize certain entities (such as low income households, or regions with declining industries) either because those entities receive the highest utility impacts, or because those entities are intrinsically more valued, irrespective of the impact of the allocation. This distinction has deep implications for understanding and designing optimal policies [21, 5]. In particular, all members of society may agree on a ranking of who benefits most along some objective metric, but may disagree on the welfare weights to assign to different entities.

This paper develops a method to infer the preferences that underlie observed allocation decisions. Our method separates differential treatment effects (who benefits the most from a program) from implied welfare weights (what types of entities are prioritized) and weights on different outcomes (how different impacts are valued). Our approach relies on recent innovations in machine learning that make it possible to estimate the heterogenous treatment effects under random assignment [17, 27]. We demonstrate how these advances can be used to better understand and articulate the allocation of programs. Our approach allows us to pose counterfactual welfare weights and valuations of outcomes to produce different allocations, and quantify the welfare impacts of these adjustments.

We consider a common form of decision, an allocation based on a score or ranking. These may be poverty scores in the case of welfare programs, or explicit rankings in the case of applicants for small business grants or college admission. This ranking implies a system of inequalities between the contributions of different entities to welfare. We use this system of inequalities, and a simple and general model of welfare, to estimate the implied value on different welfare outcomes (estimated using modern methods for heterogenous effects) and different entities (based on observed characteristics), using ordinal logit. Our method can also be used if an observer had only binary information on eligibility, though in that setting it will be less informative.

Intuitively, if a decisionmaker allocates to one type of applicant who benefits little from the allocation over a different type who would benefit greatly, that suggests the decisionmaker places higher welfare weight on the first type. Or, if a decisionmaker consistently allocates to applicants whose health improves from the allocation — instead of applicants whose consumption increases — that implies the decisionmaker highly values health.

We apply the method to the case of PROGRESA, one of the world's largest (and best-studied) anti-poverty programs, which provided cash transfers to eligible households in Mexico.[1] We first estimate the heterogenous treatment effects of the program, exploiting randomized variation in eligibility, using causal forests [27]. Results indicate treatment effect heterogeneity, such that younger and wealther households benefit most from the program. This is similar to prior work based on linear models [6]. Next, we use ordinal logit to identify the preferences implied by the ranking between households. These results indicate that the program prioritizes indigenous households, poor households, and households with children. Finally, we evaluate the counterfactual allocations that would result from alternate welfare weights. For instance, we can show what allocations *should have occurred* had the government placed higher priority on certain types of impacts (e.g., health vs. education) or certain types of households (e.g., lower-income or indigenous).

This approach makes it possible to invert the discussion about allocative programs. Rather than debate the means of the policy (who is eligible, how large are the benefits), this framework makes it possible to debate the ends (how much do we value health, education, or consumption? By how much should lower-income families be prioritized?). The framework can be applied in numerous settings where decisionmakers allocate scarce resources and heterogeneous treatment effects can be estimated. In particular, it naturally applies to the debate about universal basic income versus targeting transfers towards particular households [12].

The approach has two caveats. First, it requires defining values precisely: the implied weights may depend on what outcomes and characteristics are allowed to enter the objective function. This definition of what may be valued is a substantive decision. Second, it requires a large enough dataset to both measure heterogeneous treatment effects, and the implied welfare parameters. These datasets are increasingly becoming available, particularly in settings with digital experimentation.

Related Literature

This paper contributes to a large economics literature on optimal targeting and taxation [21, 4, 10], and especially work focused on targeting in developing countries [2, 11]. It can be viewed as a response to [24], which finds that targeting poverty directly may not be sufficient for impact, and suggests that it may be better to target based on desired outcomes. Relatedly, [28] considers the theoretical problem of allocating resources given heterogeneous aid agency preferences over individuals, and describes allocation queues as a solution to a combinatorial problem. It is related to a recently expanding public finance literature on welfare weights. [13] infers the weight on different households implied by a tax schedule based on the distortions required to transfer them resources. [25] generalize welfare weights to reconcile popular notions of fairness with optimal tax theory. Our paper shows how similar welfare questions can be raised across a broad set of domains where heterogeneous treatment effects can be estimated.

Our efforts relate broadly to recent work on fairness in machine learning [7, 3]. Within this subfield, several papers have studied the social welfare implications of algorithmic decisions, and how social welfare concerns relate to different notions of fairness [9, 14, 20, 19, 1]. Most directly related, [22] discuss how different constraints to targeting can impact efficiency and fairness. Our approach is distinct, however, in that we show how using machine learning tools can be used to better characterize and audit the implied priorities of a program, as revealed in the program's observed allocation. We hope that by providing increased visibility into these revealed preferences, future policies can be better aligned with stated preferences and explicit policy objectives.

[1] PROGRESA conditioned payouts on certain actions, but we treat the program as an unconditional transfer. For simplicity, we assume PROGRESA did not have differential spillover benefits on different households.

2 Model

We consider a decisionmaker choosing how to allocate treatment among N entities, which could be, for example, individuals, households, firms, or regions. For convenience, we refer to entities as households. The decisionmaker determines a ranking z_i for each household i, which will ultimately be used to select a treatment status $T_i \in \{0, 1\}$. Household i has characteristics $\mathbf{x}_i$.

We assume that the ranking results from some implicit welfare function:

$$S = \sum_i S_i$$

$$S_i = \mu(\mathbf{x}_i) \cdot v(T_i, \mathbf{x}_i)$$

where $\mu(\mathbf{x}_i)$ represents the welfare weight of a household with characteristics $\mathbf{x}_i$, and $v(\cdot, \cdot)$ represents the decisionmaker's implied base valuation of a household's utility.

The utility valuation may be a linear combination of multiple outcome measures (such as consumption, health, and education):

$$v(T_i, \mathbf{x}_i) = g_0(T_i, \mathbf{x}_i) + \sum_{j>0} \lambda_j g_j(T_i, \mathbf{x}_i) + C \cdot T_i$$

where $g_j(\cdot, \cdot)$ represents outcome j and λ_j represents the relative value, or 'impact weight' of outcome j relative to the numeraire ($j = 0$). C is a constant representing the intrinsic value of providing the program, even absent impact.

The policymaker may place more or less weight on households with different characteristics, with additive welfare weights $\boldsymbol{\omega}$:

$$\mu(\mathbf{x}_i) = 1 + \boldsymbol{\omega} \cdot \mathbf{x}_i$$

We take as given that we have an experimental design that has recovered the predicted effect of treatment on each outcome j, which may be heterogeneous as a function of covariates $\mathbf{x}_i$:

$$\hat{\Delta g_j}(\mathbf{x}_i) \coloneqq g_j(1, \mathbf{x}_i) - g_j(0, \mathbf{x}_i)$$

The predicted welfare impact of treating household i is then:

$$\Delta S_i = (1 + \boldsymbol{\omega} \cdot \mathbf{x}_i) \cdot (\hat{\Delta g_0}(\mathbf{x}_i) + \sum_{j>0} \lambda_j \hat{\Delta g_j}(\mathbf{x}_i) + C)$$

The decisionmaker ranks each household by its contribution to welfare:

$$z_i = f(\Delta S_i + \epsilon_i) \tag{1}$$

where ϵ_i is measurement error that is iid and mean zero, and f is a weakly increasing transformation, which preserves the priority order of who receives treatment but not the intensity of preferences.

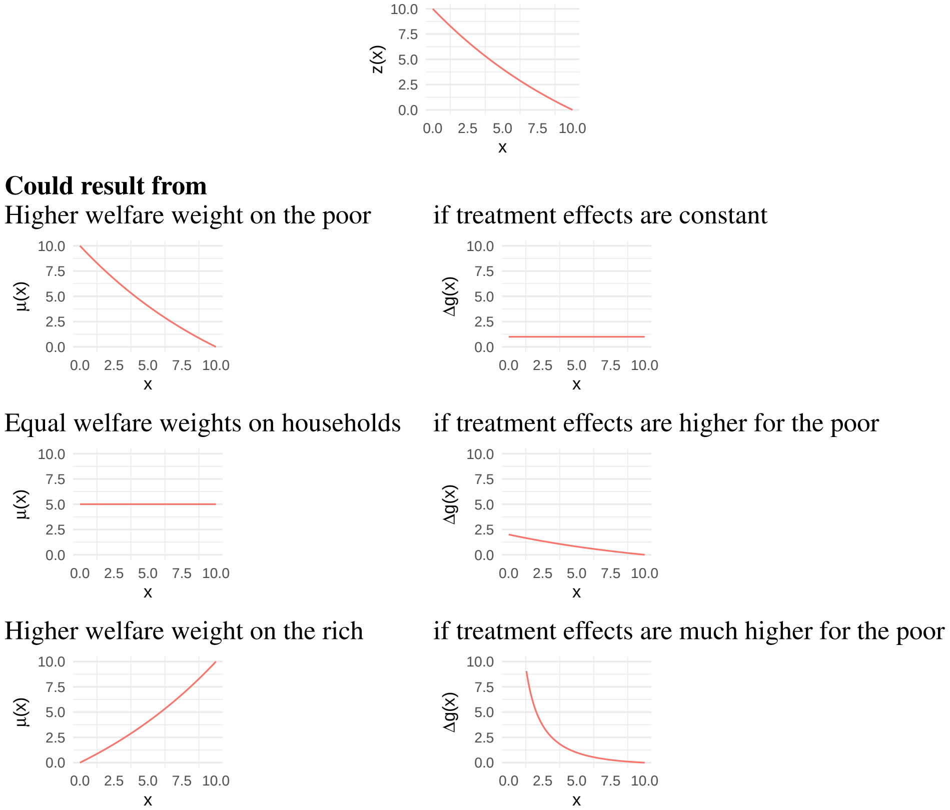
3 Intuition

To demonstrate the intuition behind our method, we consider a simple example in Figure 1. Consider the case of a single outcome and one dimension of heterogeneity, x, which corresponds with consumption. A decisionmaker allocates a program by ordering households by the function $z(x)$, prioritizing poor households. As shown in Figure 1, the same allocation rule could result from higher welfare weights on the poor, equal welfare weights, or higher welfare weights on the rich, depending on how treatment effects vary with x.

The next section demonstrates how to empirically recover welfare and impact weights from data in when there are multiple dimensions of heterogeneity and multiple outcomes of interest.

Figure 1: Intuitive Example

An allocation rule that prefers the poor (low x)...



Could result from

Higher welfare weight on the poor

if treatment effects are constant

Equal welfare weights on households

if treatment effects are higher for the poor

Higher welfare weight on the rich

if treatment effects are much higher for the poor

4 Estimation

If we observe how a decisionmaker allocates treatment ($\boldsymbol{z}$) and its effects ($\Delta\hat{g}_j(\mathbf{x}_i)$), what can we infer about the decisionmaker's preferences ($\boldsymbol{\omega}$, C, $\boldsymbol{\lambda}$)?

If the decisionmaker prioritizes household i over i' ($z_i > z_{i'}$), Equation 1 suggests we must have:

$$(1+\boldsymbol{\omega}\cdot\mathbf{x}_i)\cdot(\widehat{\Delta g}_0(\mathbf{x}_i)+\sum_{j>0}\lambda_j\widehat{\Delta g}_j(\mathbf{x}_i)+C)+\epsilon_i > (1+\boldsymbol{\omega}\cdot\mathbf{x}_{i'})\cdot(\widehat{\Delta g}_0(\mathbf{x}_{i'})+\sum_{j>0}\lambda_j\widehat{\Delta g}_j(\mathbf{x}_{i'})+C)+\epsilon_{i'}$$

If the decisionmaker's error is Gumbel: $\epsilon_i \sim \sigma \cdot EV(1)$, then the problem can be modeled with an ordinal logit likelihood. We use maximum likelihood to estimate the parameters $\{\boldsymbol{\omega}, \boldsymbol{\lambda}, C, \sigma\}$ that best match the observed data $\{\boldsymbol{z}, \mathbf{x}, \{\widehat{\Delta g}_j(\mathbf{x}_i)\}_j\}$. This method can also be used if one observes not a full ranking, but simply binary assignment to treatment ($T_i \in \{0,1\}$), which corresponds to a ranking with two levels.

5 Empirical Example

We demonstrate our method on the Mexican PROGRESA conditional cash transfer (CCT) program. First implemented by the Mexican federal government in 1997, PROGRESA was the inspiration and model for a number of similar conditional cash transfer programs across Latin America. It was designed to improve the well being of poor families, by offering monthly transfers of 90 pesos were offered to eligible mothers conditional on regular doctor's visits and/or regular school attendance. The vast majority, roughly 99%, of enrolled people met these conditions [26].[2]

5.1 Data

We use data from two household surveys prior to treatment (1996 and May 1998), and one household survey after treatment (November 1999). These surveys asked about household demographics, socioeconomic characteristics, health care utilization, and educational attendance. We evaluate endline outcomes reported in November 1999. These data contain information on 14,949 households over the entire experiment period.

5.2 Outcomes

We focus on three outcomes of interest:

- **Per-capita consumption**
- **Health status of young children**: average number of sick days among children 0-5 years old in the household
- **School attendance**: average number of school days missed among children 6-16 years old in the household

The survey asked for these responses in the previous month. In the terminology of our theory, these outcomes are our $g_j(x_i)$, where $J = 3$; we treat per-capita consumption as our numeraire. All of these are plausibly target outcomes that a social planner or government might prioritize when designing a welfare program, although only the latter two outcomes were explicitly prioritized by the Mexican federal government as part of the official goals of the program. Previous studies have estimated significant treatment impacts of PROGRESA for all three of these outcomes using the same survey data that we explore here [15, 8, 26, 6].

We impute treatment effects of 0 for households without children in the relevant age bucket for schooling and health outcomes, respectively (children aged 5 or below for the health intervention, and children aged 6-16 for the schooling intervention).

5.3 Estimates

5.3.1 Heteogeneity in treatment effects.

We first estimate heterogeneous treatment effect estimates $\Delta\hat{g}_j(\mathbf{x}_i)$ (estimated using causal forest). On average over our sample, PROGRESA increased household monthly consumption by 14 pesos, to have reduced the number of sick days per child by 0.08, and reduced the number of school days missed per child by 0.03. However, those treatment effects are heterogeneous, as summarized in Figure 2. Income and household head age are the most important covariates in explaining these differences (see Appendix Table A1).

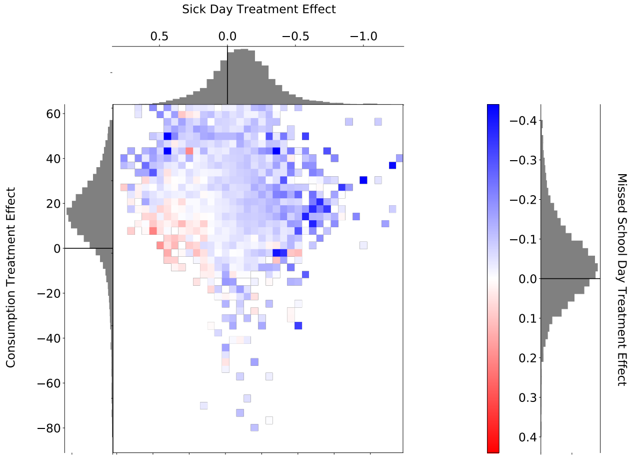
5.3.2 Implied policymaker preferences

Next, we use these estimated heterogeneous treatment effects and the household poverty scores $\boldsymbol{z}$, to estimate welfare weights and impact weights. We allow welfare weights $\mu(\mathbf{x}_i)$ to vary over the size of households, the indigenous status of the household head, the level of income in 1997, the number of adults aged 17 or above, and the number of children less than or equal to 5 years old, as well as the number of children 6 to 16 years old. We estimate parameters using ordinal logit. Standard errors are computed using a two-step bootstrap procedure that accounts for uncertainty in both treatment effects and preference parameters.[3]

[2] For a more detailed treatment of PROGRESA and its background, see [8], [15], and [26].

[3] Observations are drawn with replacement before estimation of the treatment effects and the welfare and impact weights. Treatment effects are then estimated from these bootstrapped samples, and welfare and impact weights estimated from these bootstrapped treatment effect estimates; the standard errors reported are the standard deviation across bootstrapped welfare and impact weight estimates.

Figure 2: Distribution of Estimated Treatment Effects



Notes: Joint and marginal distributions of estimated treatment effects of PROGRESA conditional cash transfer on schooling, health, and consumption, estimated using the causal forest method. Note that missed school days and sick days are inferred to be "bads", according to our estimated weights, and so higher negative values for these treatment effects are associated with higher social utility. Note also that we impute 0 for households without children in the relevant age range for health and schooling treatment effects; the above graphs show only TEs for households for which these TEs are defined.

Table 1 reports estimates. Each column represents a different targeting policy; column 1 presents estimates from the household poverty score used to allocate eligibility in our sample. We report implied decisionmaker preferences: the first block of rows shows the implied welfare weights ($\mu(\mathbf{x}_i)$), and the second block shows implied impact weights ($\boldsymbol{\lambda}$ and C) and the standard deviation of the error term (σ). The third block reports the ranking $\boldsymbol{z}$ decomposed into covariates, as described by a linear regression. The fourth block presents counterfactual average outcomes in 1999 if the same number of households were selected under that targeting policy.

We find that allocations are consistent with welfare weights that rank households 13.3 percentiles higher if indigenous, 7.7 percentiles lower for each standard deviation increase in household income, 21.1 percentiles higher for each additional small child (ages 5 and lower) in the household, 15.1 percentiles higher for each additional child aged 6-16, and 20.3 percentiles lower for additional adult. The coefficients on income, children and adults are statistically significantly different from 0 at a 5% level. The welfare weight on indigenous status is estimated slightly less precisely, with a T stat of 1.17.[4]

Most of the implied value of the program comes from simply providing the program, independent from its effect on outcomes (the constant term C). C accounts for 98% of the welfare gain for the median household.

The weights on individual outcomes are measured less precisely. 95 confidence intervals suggest that the Mexican government's initial allocation rule implies a value of each missed school day among children below 685 pesos of consumption, and a value of a sick days among young children below 690.40 pesos, but confidence intervals span zero. These valuations can be compared against other estimates of the value of education and health. Based on a review of multiple studies, [23] suggest a 9% average return to a year of schooling. Assuming that these gains accrue once the child is working age, with a lifetime of 40 years of work and a discount rate of 3%, this corresponds to lifetime present-discounted earnings of 2143.35 pesos or \$214.34 per missed year of school.[5] Our 95% confidence intervals suggest a value per missed year of school below \$1933.12. WHO recommendations which consider health interventions to be a 'best buy' if they can save a DALY

[4] We compute the average percentile change by first computing how each household's projected ranking would shift, given different covariates, and then taking the median change over all households.

[5] Mean monthly consumption is 234.96 pesos. We compute $\sum_{y=16}^{16+40}(0.09*234.96*12)*(0.97)^{16-7.5+y} =$ 2143.35, where 7.5 is the average child age in our sample.

Table 1: Allocation Rules

	Actual	Counterfactual							
	Household Poverty Score	Empirical Impact Weights		Technocratic Impact Weights		Only Value Consumption Impacts	Only Value Health Impacts	Only Value Education Impacts	2003 Household Poverty Score
		Equal Welfare Weights	Prefer Poorest	Empirical Welfare Weights	Equal Welfare Weights	Empirical Welfare Weights	Empirical Welfare Weights	Empirical Welfare Weights	
Welfare weights $\mu(\mathbf{x}_i)$									
$\omega_{Indigenous}$	0.116 (0.099)	1	0	0.116	1	0.116	0.116	0.116	0.191 (0.535)
$\omega_{log(1997Inc.)}$	-0.092 (0.043)	1	-1	-0.092	1	-0.092	-0.092	-0.092	-0.01 (0.018)
$\omega_{NumAdults}$	-0.196 (0.086)	1	0	-0.196	1	-0.196	-0.196	-0.196	-0.09 (0.286)
$\omega_{NumChild<=5y.o.}$	0.209 (0.096)	1	0	0.209	1	0.209	0.209	0.209	0.146 (0.314)
$\omega_{NumChild6to16y.o.}$	0.143 (0.073)	1	0	0.143	1	0.143	0.143	0.143	0.101 (0.245)
Impact weights $\boldsymbol{\lambda}$									
$\lambda_{Consumption}$ (pesos)	1.0	1.0	1.0	1.0	1.0	1.0	0	0	1.0
λ_{Health} (pesos per child sick day)	11.749 (350.648)	11.749	11.749	-35.714	-35.714	0	11.749	0	253.324 (387.008)
$\lambda_{Schooling}$ (pesos per missed school day)	-16.838 (368.354)	-16.838	-16.838	-96.507	-96.507	0	0	-16.838	206.78 (729.306)
C (pesos regardless of impact)	205.394 (5079.108)	205.394	205.394	205.394	205.394	0	0	0	1946.606 (153094.424)
σ	0.006 (0.002)	.	.	.	.	.	.	.	0.0015 (0.0006)
Score Function (Linear regression coeffs.) $z(\mathbf{x}_i)$									
$\alpha_{Indigenous}$	2.224	2.935	-0.654	2.411	2.612	2.304	2.672	2.729	9.476
$\alpha_{log(1997Inc.)}$	-2.867	2.088	-5.444	-3.176	2.492	-2.897	-2.097	-2.186	-0.482
$\alpha_{NumAdults}$	-5.265	1.795	0.011	-4.945	1.549	-5.264	-4.646	-4.606	-4.528
$\alpha_{NumChild<=5y.o.}$	5.348	2.287	-0.231	4.889	2.648	5.342	4.963	4.892	7.3
$\alpha_{NumChild6to16y.o.}$	3.638	2.145	-0.145	3.36	2.335	3.662	3.369	3.33	5.067
Pred. outcomes under alt. treatment assignment									
Monthly consumption, 1999 (pesos)	223.28746	230.16148	220.45564	223.21529	230.18798	223.30823	225.24285	225.22214	227.49874
Monthly sick days, 1999 (per child)	1.09446	1.12393	1.11583	1.10535	1.11456	1.09619	1.09257	1.09512	1.09557
Monthly missed school days, 1999 (per school-age child)	0.27933	0.27519	0.2899	0.28364	0.27221	0.2786	0.27709	0.27803	0.27387
N	13438	13438	13438	13438	13438	13438	13438	13438	13438

for \$100 [18], as well as the revealed preference valuations of [16] of \$23.68 per DALY for Kenyan households, based on how far they are willing to walk for clean water. If we count each day a child is sick as a day lost, so that a full year of sickness would represent a disability adjusted life year (DALYs), then the government's allocation is consistent with a value per DALY below \$1917.97.[6] Achieving more precise estimates of impact weights may require a larger sample.

5.4 Counterfactuals

We compare the estimates from the government allocation to counterfactual allocations in the remaining columns of Table 1.

Alternate welfare weights. When welfare weights are set equal across households (column 2), the resulting score puts positive weight on log income and puts relatively less weight on households with more small children, and positive weight on households with more adults. When welfare weights rank households solely by log income (column 3), the resulting score deprioritizes households with more small children and also deprioritizes indigenous households.

Technocratic impact weights. In columns 4-5 of Table 1, we keep the original welfare weights but assume technocratic impact weights, as might be input from external valuations. We do not intend to take a stand on these valuations, so these results should be viewed as speculative. We demonstrate results assuming valuations of 1000 pesos (\$100) per DALY and 2143.35 pesos per missed school day. The z' ranking implied by our assumed weights is quite similar to the original. By contrast, changing the household covariate welfare weights to an equal weighting across households so that $\mu(\mathbf{x}_i) \equiv 1$ leads to positive weight on income and on household number of adults.

Focus on different outcomes. In columns 6-8 of Table 1, we present alternative specifications that alter impact weights to be entirely determined by one of the three outcomes. When the impact weights are determined entirely by health effects, z' prioritizes non-indigenous households and also households with fewer children; when the impact weights are determined entirely by consumption outcomes, z' upweights households where the head is indigenous and households with more children; when impact weights are determined by schooling outcomes, z' upweights households with higher income and fewer small children. In all three of these settings, smaller households (in terms of members of all ages) are given higher priority.

An alternative government scoring rule. In 2003, the Mexican government expanded PROGRESA, and updated their poverty score. In column 9, we find that the welfare weights implied by this rule are largely similar to the welfare weights from the 1997 original ranking, but with less negative weight on income and on households with fewer children, and higher weight on indigenous status. On average, the government would rank households 36.8 percentiles higher if indigenous, 1.4 percentiles lower for each standard deviation increase in log household income, 16.5 percentiles lower for each additional adult, 18.4 percentiles higher for each additional child aged 6-16, and 25.2 percentiles higher for additional child aged 5 and lower. These covariates are estimated with considerably less precision, however, and none are statistically significantly different from 0. The impact weights are imprecisely estimated; we find the valuation of a missed day of school has a 95% confidence interval below 1607 pesos, and the valuation of a young child sick day below 489.73 pesos.

6 Conclusion

While analysts reason about primitives of utility and welfare weights, policy discussions commonly revolve instead around the mechanics of implementation. This paper demonstrates how heterogeneous treatment effect estimates can be used to bridge between these two conceptions. This framework could be used in several ways. First, it could be used to characterize existing allocations, to provide an indication of the preferences they imply. This, in turn, provides an ex-post auditing mechanism that can help hold decisionmakers accountable for past transfers – and in particular, to evaluate whether the implemented allocation reflects the stated goals of the policy. Perhaps most importantly, this approach can be used to amend existing policies and guide future allocations. In particular, it can demonstrate how different priorities over welfare outcomes and population subgroups would produce different allocations, and quantifies the welfare impacts of these adjustments.

[6] As a rough conversion between sick days and DALYs, we multiply the value of each sick day by the 28 days asked over the survey to convert the valuation of sick days to the valuation of a 'sick year.'

References

[1] Abebe, R., Kleinberg, J., and Weinberg, S. M. Subsidy Allocations in the Presence of Income Shocks. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 05 (Apr. 2020), 7032–7039. Number: 05.

[2] Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., and Tobias, J. Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review 102*, 4 (June 2012), 1206–1240.

[3] Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2018.

[4] Barr, N. *Economics of the welfare state*. Oxford university press, 2012.

[5] Coate, S., and Morris, S. On the Form of Transfers to Special Interests. *Journal of Political Economy 103*, 6 (Dec. 1995), 1210–1235.

[6] Djebbari, H., and Smith, J. Heterogeneous impacts in PROGRESA. *Journal of Econometrics 145*, 1 (July 2008), 64–80.

[7] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (2012), ACM, pp. 214–226.

[8] Emmanuel Skoufias, V. D. M. Conditional Cash Transfers, Adult Work Incentives, and Poverty. *Journal of Development Studies 44*, 7 (2008), 935–960.

[9] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway Feedback Loops in Predictive Policing. *arXiv:1706.09847 [cs, stat]* (June 2017). arXiv: 1706.09847.

[10] Fleurbaey, M., and Maniquet, F. Optimal income taxation theory and principles of fairness. *Journal of Economic Literature 56*, 3 (2018), 1029–79.

[11] Hanna, R., and Olken, B. A. Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries. *Journal of Economic Perspectives 32*, 4 (Nov. 2018), 201–226.

[12] Hanna, R., and Olken, B. A. Who should receive anti-poverty programs? Universal basic incomes vs. targeted transfers in developing countries. *Journal of Economic Perspectives* (2018).

[13] Hendren, N. Efficient Welfare Weights. Working Paper 20351, National Bureau of Economic Research, 2019.

[14] Hu, L., and Chen, Y. Welfare and Distributional Impacts of Fair Classification. *arXiv:1807.01134 [cs, stat]* (July 2018). arXiv: 1807.01134.

[15] John Hoddinott, E. S. The Impact of PROGRESA on Food Consumption. *Economic Development and Cultural Change 53*, 1 (2004), 37–61.

[16] Kremer, M., Leino, J., Miguel, E., and Zwane, A. P. Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions. *The Quarterly Journal of Economics 126*, 1 (Feb. 2011), 145–205.

[17] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences 116*, 10 (Mar. 2019), 4156–4165.

[18] Laxminarayan, R., Chow, J., and Shahid-Salles, S. A. *Intervention Cost-Effectiveness: Overview of Main Messages*. The International Bank for Reconstruction and Development / The World Bank, 2006.

[19] LIU, L. T., DEAN, S., ROLF, E., SIMCHOWITZ, M., AND HARDT, M. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning* (Stockholm, Sweden, 2018), vol. 80 of *Proceedings of Machine Learning Research*, pp. 3156–3164.

[20] MOUZANNAR, H., OHANNESSIAN, M. I., AND SREBRO, N. From Fair Decision Making to Social Equality. *arXiv:1812.02952 [cs, stat]* (Dec. 2018). arXiv: 1812.02952.

[21] NICHOLS, A. L., AND ZECKHAUSER, R. J. Targeting Transfers through Restrictions on Recipients. *The American Economic Review 72*, 2 (1982), 372–377.

[22] NORIEGA, A., GARCIA-BULLE, B., TEJERINA, L., AND PENTLAND, A. Algorithmic Fairness and Efficiency in Targeting Social Welfare Programs at Scale. *Bloomberg Data for Good Exchange Conference* (2018).

[23] PSACHAROPOULOS, G., AND PATRINOS, H. A. Returns to investment in education.

[24] RAVALLION, M. How Relevant Is Targeting to the Success of an Antipoverty Program? *The World Bank Research Observer 24*, 2 (2009), 205–231.

[25] SAEZ, E., AND STANTCHEVA, S. Generalized Social Marginal Welfare Weights for Optimal Tax Theory. *American Economic Review 106*, 1 (Jan. 2016), 24–45.

[26] SIMONE BOYCE, P. G. An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico. vol. 85, Royal Economic Society.

[27] WAGER, S., AND ATHEY, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association 113*, 523 (July 2018), 1228–1242.

[28] WANG, F. The Optimal Allocation of Resources Among Heterogeneous Individuals. *Available at SSRN* (2020).

A Notes for NeurIPS ML for Economic Policy

This paper provides a method to allow people to inspect the values underlying policies, and policymakers to create policies with values in mind.

We hope that this method will improve the alignment of policies and values.

This paper has not been published or first made available before January 1, 2017, and is original work.

Table A1: Feature Importance Estimates: Causal Forest

	Consumption (Monthly avg. per person, in pesos)	Health (Avg. sick days per child)	Schooling (Avg. days of missed school per child)
head age	0.351	0.181	0.216
log household income ’97	0.204	0.192	0.329
household size	0.095	0.236	0.14
head education	0.19	0.151	0.082
num children less than 2 yrs old	0.025	0.045	0.086
num children 3 to 5 yrs	0.009	0.057	0.034
head agricultural worker	0.016	0.048	0.033
male head of household	0.038	0.021	0.011
num women at least 55 yrs	0.022	0.02	0.017
num women 20 to 34 yrs	0.018	0.015	0.019
num children 6 to 10 yrs	0.005	0.022	0.016
num men at least 55 yrs	0.018	0.005	0.012
num boys 15 to 19 yrs	0.008	0.008	0.005
num boys 11 to 14 yrs	0.0	0.0	0.0
num men 20 to 34 yrs	0.0	0.0	0.0
num women 35 to 54 yrs	0.0	0.0	0.0
num girls 15 to 19 yrs	0.0	0.0	0.0
head indigenous	0.0	0.0	0.0
num girls 11 to 14 yrs	0.0	0.0	0.0
num men 35 to 54 yrs	0.0	0.0	0.0
N	13438	8769	9871